Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art

Jung-ran Park
and
Andrew Brenza

ABSTRACT

Assessment of the current landscape of semi-automatic metadata generation tools is particularly important considering the rapid development of digital repositories and the recent explosion of big data. Utilization of semi-automatic metadata generation is critical in addressing these environmental changes and may be unavoidable in the future considering the costly and complex operation of manual metadata creation. To address such needs, this study examines the range of semi-automatic metadata generation tools (N = 39) while providing an analysis of their techniques, features, and functions. The study focuses on open-source tools that can be readily utilized in libraries and other memory institutions. The challenges and current barriers to implementation of these tools were identified. The greatest area of difficulty lies in the fact that the piecemeal development of most semi-automatic generation tools only addresses part of the issue of semi-automatic metadata generation, providing solutions to one or a few metadata elements but not the full range of elements. This indicates that significant local efforts will be required to integrate the various tools into a coherent set of a working whole. Suggestions toward such efforts are presented for future developments that may assist information professionals with incorporation of semi-automatic tools within their daily workflows.

INTRODUCTION

With the rapid increase in all types of information resources managed by libraries over the last few decades, the ability of the cataloging and metadata community to describe those resources has been severely strained. Furthermore, the reality of stagnant and decreasing library budgets has prevented the library community from addressing this issue with concomitant staffing increases. Nevertheless, the ability of libraries to make information resources accessible to their communities of users remains a central concern. Thus there is a critical need to devise efficient and cost effective ways of creating bibliographic records so that users are able to find, identify, and obtain the information resources they need.

One promising approach to managing the ever-increasing amount of information is with semi-automatic metadata generation tools. Semi-automatic metadata generation tools

Jung-ran Park (jung-ran.park@drexel.edu) is Editor, *Journal of Library Metadata*, and Associate Professor, College of Computing and Informatics, Drexel University, Philadelphia.
Andrew Brenza (apb84@drexel.edu) is Project Assistant, College of Computing and Informatics, Drexel University, Philadelphia.

concern the use of software to create metadata records with varying degrees of supervision from a human specialist.[1] In its ideal form, semi-automatic metadata generation tools are capable of extracting information from structured and unstructured information resources of all types and creating quality metadata that not only facilitate bibliographic record creation but also semantic interoperability, a critical factor for resource sharing and discovery in the networked environment. Through the use of semi-automatic metadata generation tools, the library community has the potential to address many issues related to the increase of information resources, the strain on library budget, the need to create high-quality, interoperable metadata records, and, ultimately, the effective provision of information resources to users.

There are many potential benefits to semi-automatic metadata generation. The first is scalability. Because of the quantity of information resources and the costly and time-consuming nature of manual metadata generation,[2] it is increasingly apparent that there simply are not enough information professionals available for satisfying the metadata-generation needs of the library community. Semi-automatic metadata generation, on the other hand, offers the promise of using high levels of computing power to manage large amounts of information resources. In addition to scalability, semi-automatic metadata generation also offers potential cost savings through a decrease in the time required to create effective records. Furthermore, the time savings would allow information professionals to focus on tasks that are more conceptually demanding and thus not suitable for automatic generation. Finally, because computers can perform repetitive tasks with relative consistency when compared to their human counterparts, automatic metadata generation promises the ability to create more consistent records. A potential increase in consistency of quality metadata records would, in turn, increase the potential for interoperability and thereby the accessibility of information resources in general. Thus semi-automatic metadata generation offers the potential to not only ease resource description demands on the library community but also to improve resource discovery for its users.

GOALS OF THE STUDY

Assessment of the current landscape of semi-automatic metadata generation tools is particularly important considering the fast development of digital repositories and the recent explosion of data and information. Utilization of semi-automatic metadata generation is critical to address such environmental changes and may be unavoidable in the future considering the costly and complex operation of manual metadata creation. Even though there are promising experimental studies that exploit various methods and sources for semi-automatic metadata generation,[3] a lack of studies assessing and evaluating the range of tools have been developed, implemented, or improved. To address such needs, this study aims to examine the current landscape of semi-automatic metadata generation tools while providing an evaluative analysis of their techniques, features, and functions. The study primarily focuses on open-source tools that can be readily utilized in libraries and other memory institutions. The study also highlights some of the challenges still facing the continued development of semi-automatic tools and the current barriers

doi: 10.6017/ital.v34i3.5889

to their incorporation into the daily workflows for information organization and management. Future directions for the further development of tools are also discussed.

Toward this end, a critical review of the literature in relation to semi-automatic metadata generation tools published from 2004 to 2014 was conducted. Databases such as Library and Information Sciences Abstracts and Library, Information Science and Technology Abstracts were searched and germane articles identified through review of titles and abstracts. Because the problem of creating viable tools for the reliable automatic generation of metadata is a not a problem limited to the library and information science professions,[4] database searches were expanded to include those databases pertinent to the computing science, including Proquest Computing, Academic Search Premier, and Applied Science and Technology. Keywords, such as "automatic metadata generation," "metadata extraction," "metadata tools," and "text mining," including their stems, were used to explore the databases. In addition to keyword searching, relevant articles were also identified within the reference sections of articles already deemed pertinent to the focus of the survey as well as through the expansion of results lists through the application of relevant subject terms applied to pertinent articles. To ensure that the latest, most reliable developments in automatic metadata were reviewed, various filters, such as date range and peer-review, were employed. Once tools were identified, their capabilities were tested (when possible), their features were noted, and overarching developments were determined.

The remainder of the article provides an overview of the primary techniques developed for the semi-automatic generation of metadata and a review of the open-source metadata generation tools that employ them. The challenges and current barriers to semi-automatic metadata tool implementation are described as well as suggestions for future developments that may assist information professionals with integration of semi-automatic tools within the daily workflow of technical services departments.

Current Techniques for the Automatic Generation of Metadata

As opposed to manual metadata generation, semi-automatic metadata generation relies on machine methods to assist with or to complete the metadata-creation process. Greenberg distinguished between two methods of automatic metadata generation: metadata extraction and metadata harvesting.[5] Metadata extraction in general employs automatic indexing and information retrieval techniques to generate structured metadata using the original content of resources. On the other hand, metadata harvesting concerns a technique to automatically gather metadata from individual repositories in which metadata has been produced by semi-automatic or manual approaches. The harvested metadata can be stored in a central repository for future resource retrieval.

Within this dichotomy of extraction methods, there are several other more specific techniques that researchers have developed for the semi-automatic generation of metadata. Polfreman et al. identified an additional six techniques that have been developed over the years: meta-tag harvesting, content extraction, automatic indexing, text and data mining, extrinsic data auto

generation, and social tagging.[6] Although the last technique is not properly a semi-automatic metadata generation technique because it is used to generate metadata with a minimum of intervention required by metadata professionals, it can be viewed as a possible mode to streamline the metadata creation process.

Both Greenberg and Polfreman provide comprehensive, high-level characterizations of the techniques employed in current semi-automatic metadata generation tools. However, an evaluation of these techniques within the context of a broad survey of the tools themselves and a comprehensive enumeration of currently available tools are not addressed. Thus, although these techniques will be examined for the remainder of this section, they serve simply as a framework through which this study provides a current and comprehensive analysis of the tools available for use today. Each section provides an overview of the relevant technique, a discussion of the most current research related to it, and the tools that employ that technique.

The tables included in each section provide lists of the semi-automatic metadata generation tools (N = 39) evaluated in the course of this survey. The information presented in the tables is designed to provide a characterization of each tool: its name, its online location, the technique(s) used to generate metadata, and a brief description of the tool's functions and features. Only those tools that are currently available for download or for use as web services at the time of this writing are included. Furthermore, the listed tools have not been strictly limited to metadata-generation applications but also include some content management system software (CMSS) as these generally provide some form of semi-automatic metadata extraction. Typically, CMSS are capable of extracting technical metadata as well as data that can found in the meta-tags of information resources, such as the file name, and using that information as the title of a record.

Meta-Tag Extraction

Meta-tag extraction is a computing process whereby values for metadata fields are identified and populated through an examination of metadata tags within or attached to a document. In other words, it is a form of metadata harvesting and, possibly, conversion of that metadata into other formats. MarcEdit, the most widely used semi-automatic tool for the generation of metadata in US libraries,[7] is an example of this technique. MarcEdit essentially harvests metadata from Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant records and offers the user the opportunity to convert those records to a variety of formats, including MAchine-Readable Cataloging (MARC), MAchine-Readable Cataloguing in XML (MARC XML), Metadata Object Description Schema (MODS), and Encoded Archival Description (EAD). It also offers the capabilities of converting records from any of the supported formats to any of the other supported formats.

Other examples of this technique are the web services Editor-Converter Dublin Core Metadata and Firefox Dublin Core Viewer Extension. Both of these programs search HTML files on the web and convert information found in HTML meta-tags to Dublin Core elements. In the cases of MarcEdit

doi: 10.6017/ital.v34i3.5889

and Editor-Converter Dublin Core, users are presented with the converted information in an interface that allows the user to edit or refine the data.

Figure 1 provides an illustration of the extracted metadata of the *New York Times* homepage using Editor-Converter Dublin Core, while figure 2 offers an illustration of the editor that this web service provides.

Figure 1. Screenshot of Extracted Dublin Core Metadata Using Editor-Converter Dublin Core.

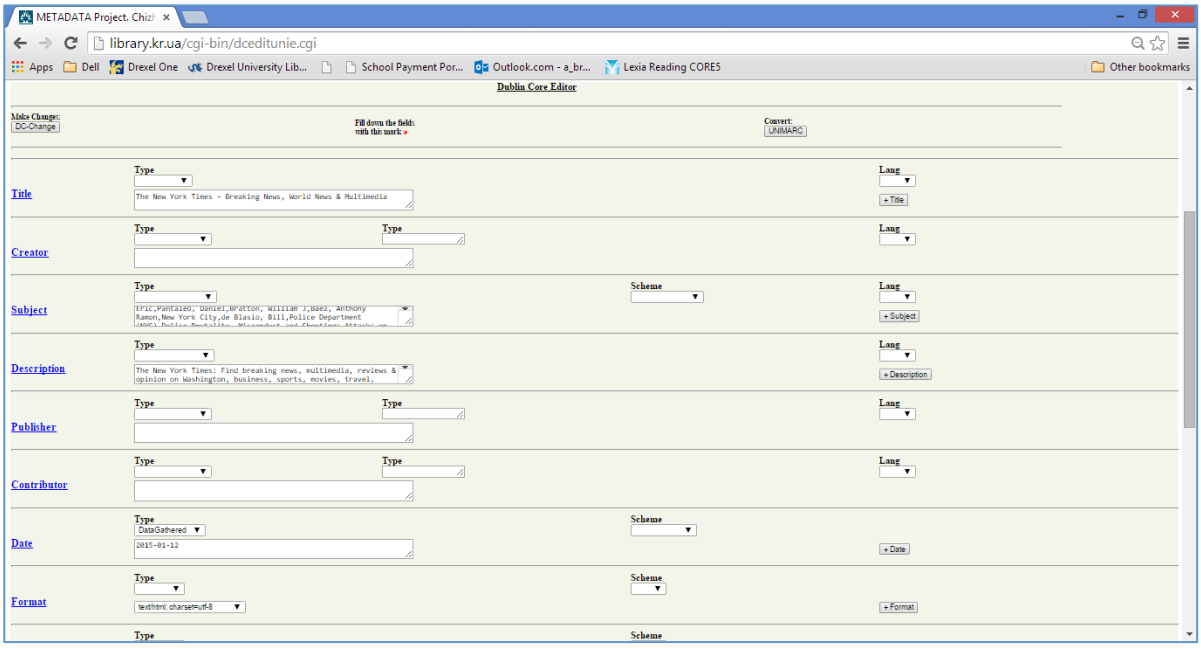

Figure 2. Screenshot of Editor-Converter Dublin Core Editing Tool (only eight of the sixteen fields are visible in this screenshot).

Perhaps the biggest weakness to this type of tool is that it entirely depends on the quality of the metadata from which the programs harvest. This can be most readily seen in the above figure by the lack of values for a number of the Dublin Core fields for the *The New York Times* website. Programs that solely employ the technique of meta-tag harvesting are unable to infer values for metadata elements that are not already populated in the source.

Table 1 lists the tools that support meta-tag harvesting either as the sole technique or as one of a suite of techniques used to generate metadata from resources. Of the thirty-nine tools evaluated for this study, nineteen support meta-tag harvesting.

Tool Name	***Location***	***Techniques***	***Functions/Features***
ANVL/ERC Kernel Metadata Conversion Toolkit	http://search.cpan.org/~jak/File-ANVL/anvl	meta-tag harvester	A utility that can automatically convert records in the ANVL format into other formats such as XML, JSON (JavaScript Object Notation), Turtle or Plain, among others.
Apache POI – Text Extractor	http://poi.apache.org/download.html	content extractor; meta-tag harvester; extrinsic auto-generator	Apache POI provides basic text extraction for all project supported file formats. In addition to the (plain) text, Apache POI can access the metadata associated with a given file, such as title and author.
Apache Tika	http://tika.apache.org/	content extractor; meta-tag harvester; extrinsic auto-generator	Built on Apache POI, the Apache Tika toolkit detects and extracts metadata and text content from various documents.
Ariadne Harvester	http://sourceforge.net/projects/ariadnekps/files/?source=navbar	meta-tag harvester	A harvester of OAI-PMH compliant records which can be converted to various other schema such as *Learning Object Metadata (LOM)*.
BIBFRAME Tools	http://www.loc.gov/bibframe/implementation/	meta-tag harvester	BIBFRAME offers a number of tools for the conversion of MARCXML documents to BIBFRAME documents. Web service and downloadable software are both available.
Data Fountains	http://datafountains.ucr.edu/	content extractor; automatic indexer; meta-tag harvester; extrinsic auto-generator	Scans HTML documents and first extracts information contained in meta-tags. If information is unavailable in meta-tags, the program will use other techniques to assign values. Includes a focused web crawler that can target websites concerning a specific subject.

doi: 10.6017/ital.v34i3.5889

Dublin Core Meta Toolkit	http://sourceforge.net/projects/dcmetatoolkit/files/?source=navbar	meta-tag harvester	Transforms data collected via different methods into Dublin Core (DC) compatible metadata.
Dspace	http://www.dspace.org/	meta-tag harvester; extrinsic auto-generator; social tagging	Automatically extracts technical information regarding file format and size. Can also extract some information from meta-tags.
Editor-Converter Dublin Core Metadata	http://www.library.kr.ua/dc/dceditunie.html	meta-tag harvester; extrinsic auto-generator	Scans HTML documents, harvesting metadata from tags and converting them to DC.
Embedded Metadata Extraction Tool (EMET)	http://www.artstor.org/global/g-html/download-emet-public.html	content extractor; meta-tag harvester; extrinsic auto-generator	EMET is a tool designed to extract metadata embedded in JPEG and TIFF files.
Firefox Dublin Core Viewer Extension	http://www.splintered.co.uk/experiments/73/	meta-tag harvester; extrinsic auto-generator	Scans HTML documents, harvesting metadata from tags and displaying them in Dublin Core.
MarcEdit	http://marcedit.reeset.net/	meta-tag harvester	Harvests OAI-PMH compliant data and converts it to various formats including DC and MARC.
Metatag Extractor Software	http://meta-tag-extractor.software.informer.com/	meta-tag harvester	Permits customizable extraction features, harvesting meta-tags as well as contact information from websites.
My Meta Maker	http://old.isn-oldenburg.de/services/mmm/	meta-tag harvester	Can convert manually entered data into DC.
Photo RDF-Gen	http://www.webposible.com/utilidades/photo_rdf_generator_en.html	meta-tag harvester	Generates Dublin Core and Resource Description Framework (RDF) output from manually entered input.
PyMarc	https://github.com/edsu/pymarc	meta-tag harvester	Scripting tool in Python language for the batch processing of MARC records, similar to MarcEdit.
RepoMMan	http://www.hull.ac.uk/esig/repomman/index.html	meta-tag harvester; content extractor; extrinsic auto-generator	Automatically extracts various elements for documents uploaded to Fedora such as author, title, description, and key words, among others. Results are presented to user for review.
SHERPA/RoMEO	http://www.sherpa.ac.uk/romeo/api.html	meta-tag harvester	A machine-to-machine Application Program Interface (API) that permits the automatic look-up and importation of publishers and journals.
URL and Metatag Extractor	http://www.metatagextractor.com/	meta-tag harvester	Permits the targeted searching of websites and extracts URLs and meta-tags from those sites.

Table 1. Semi-Automatic Tools that Support Meta-Tag Harvesting.

Content Extraction

Content extraction is a form of metadata extraction whereby various computing techniques are used to extract information from the information resource itself. In other words, these techniques do not rely on the identification of relevant meta-tags for the population of metadata values. An example of this technique is the Kea application, a program developed at the New Zealand Digital Library that uses machine learning, term frequency-inverse document frequency (TF.IDF) and first-occurrence techniques to identify and assign key phrases from the full text of documents.[8] The major advantage of this type of technique is that the extraction of metadata can be done independently of the quality of metadata associated with any given information resource. Another example of a tool utilizing this technique is the Open Text Summarizer, an open-source program that offers the capability of reading a text and extracting important sentences to create a summary as well as to assign keywords. Figure 3 provides a screenshot of what a summarized text might look like using the Open Text Summarizer.

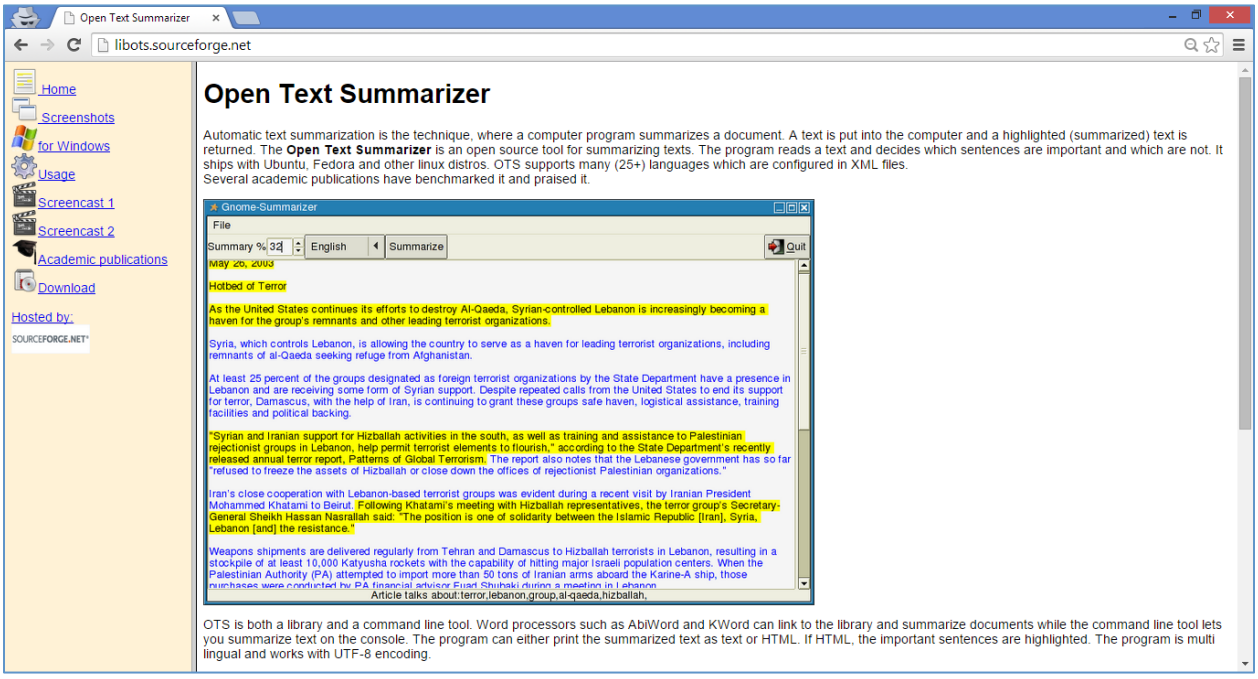

Figure 3. Open Text Summarizer: Sample Summary of Text.

Another form of this technique often relies on the predictable structure of certain types of documents to identify candidate values for metadata elements. For instance, because of the reliable format of scholarly research papers—which generally include a title, author, abstract, introduction, conclusion, and reference sections in predictable ways—this format can be exploited by machines to extract metadata values from them. Several projects have been able to exploit this technique in combination with machine learning algorithms to extract various forms of metadata.

For instance, in the Randkte project, optical character recognition software was used to scan a large quantity of legal documents from which, because of the regularity of the documents'

doi: 10.6017/ital.v34i3.5889

structure, structural metadata such as chapter, section, and page number could be extracted.[9] In contrast, the Kovacevic's project used the predictable structure of scholarly articles, converting documents from PDF to HTML files while preserving the formatting details and used classification algorithms to extract metadata regarding title, author, abstract, and keywords, among other elements.[10]

Table 2 lists the tools that support content extraction either as the sole technique or as one of a suite of techniques used to generate metadata from resources. Of the thirty-nine tools evaluated for this study, twenty tools support some form of content extraction.

Tool Name	Location	Techniques	Functions/Features
Apache POI—Text Extractor	http://poi.apache.org/download.html	content extractor; meta-tag harvester; extrinsic auto-generator	Apache POI provides basic text extraction for all project supported file formats. In addition to the (plain) text, Apache POI can access the metadata associated with a given file, such as title and author.
Apache Standol	https://stanbol.apache.org/	content extractor; automatic indexer	Extracts semantic metadata from PDF and text files. Can apply extracted terms to ontologies.
Apache Tika	http://tika.apache.org/	content extractor; meta-tag harvester; extrinsic auto-generator	Built on Apache POI, the Apache Tika toolkit detects and extracts metadata and text content from various documents.
Biblio Citation Parser	http://search.cpan.org/~mjewell/Biblio-Citation-Parser-1.10/	content extractor	A set of modules for citation parsing.
CatMDEdit	http://catmdedit.sourceforge.net/	content extractor	CatMDEdit allows the automatic creation of metadata for collections of related resources, in particular spatial series that arise as a result of the fragmentation of geometric resources into datasets of manageable size and similar scale.
CrossRef	http://www.crossref.org/SimpleTextQuery/	content extractor	This web service returns Digital Object Identifiers for inputted references.
Data Fountains	http://datafountains.ucr.edu/	content extractor; automatic indexer; meta-tag harvester; extrinsic auto-generator	Scans HTML documents and first extracts information contained in meta-tags. If information is unavailable in meta-tags, the program will use other techniques to assign values. Includes a focused web crawler that can target websites concerning a specific subject.

Embedded Metadata Extraction Tool (EMET)	http://www.artstor.org/global/g-html/download-emet-public.html	content extractor; meta-tag harvester; extrinsic auto-generator	EMET is a tool designed to extract metadata embedded in JPEG and TIFF files.
FreeCite	http://freecite.library.brown.edu/	content extractor	Free parsing tool for the extraction of reference information. Can be downloaded or used as a web service.
General Architecture for Text Engineering (GATE)	http://gate.ac.uk/overview.html	content extractor; automatic indexer;	Natural language processor and information extractor.
Kea	http://www.nzdl.org/Kea/index_old.html#download	content extractor; automatic indexer	Analyzes the full texts of resources and extracts keyphrases. Keyphrases can also be mapped to customized ontologies or controlled vocabularies for subject term assignment.
MetaGen	http://www.codeproject.com/Articles/41910/MetaGen-A-project-metadata-generator-for-Visual-St	content extractor; automatic indexer	Used to build a metadata generator for Silverlight and Desktop CLR projects, MetaGen can be used as a replacement for static reflection (expression trees), reflection (walking the stack), and various other means for deriving the name of a property, method, or field.
MetaGenerator	http://extensions.joomla.org/extensions/site-management/seo-a-metadata/meta-data/11038	content extractor	A plugin that automatically generates description and keyword meta-tags by pulling text from joomla content. With this plugin you can also control some title options and add URL meta-tags.
Ont-O-Mat	http://projects.semwebcentral.org/projects/ontomat/	content extractor	Assists user with annotation of websites that are Semantic Web-compliant. May now include a feature that automatically suggests portions of the website to annotate.
Open Text Summarizer	http://libots.sourceforge.net/	content extractor	Extracts pertinent sentences from a resource to build a free text description.

doi: 10.6017/ital.v34i3.5889

ParsCit	http://wing.comp.nus.edu.sg/parsCit/#ws	content extractor	Open-source string-parsing package for the extraction of reference information from scholarly articles.
RepoMMan	http://www.hull.ac.uk/esig/repomman/index.html	meta-tag harvester; content extractor; extrinsic auto-generator	Automatically extracts various elements for documents uploaded to Fedora such as author, title, description, and key words, among others. Results are presented to user for review.
Simple Automatic Metadata Generation Interface (SamgI)	http://hmdb.cs.kuleuven.be/amg/Download.php	content extractor; extrinsic auto-generator	A suite of tools that is able to automatically extract metadata elements such as key phrase and language from documents as well as from the context in which a document exists.
Termine	http://www.nactem.ac.uk/software/termine/	content extractor	Extracts keywords from texts through C-value analysis and Acromine, an acronym identifier and dictionary. Available as free web service for academic use.
Yahoo Content Analysis API	https://developer.yahoo.com/contentanalysis/	content extractor; automatic indexer	The Content Analysis Web Service detects entities/concepts, categories, and relationships within unstructured content. It ranks those detected entities/concepts by their overall relevance, resolves those if possible into Wikipedia pages, and annotates tags with relevant metadata.

Table 2. Semi-automatic Tools that Support Content Extraction

Automatic Indexing

In the same way as content extraction, automatic indexing involves the use of machine learning and rule-based algorithms to extract metadata values from within information resources themselves, rather than relying on the content of meta-tags applied to resources. However, this technique also involves the mapping of extracted metadata terms to controlled vocabularies such as the Library of Congress Subject Headings (LCSH), the Getty Thesaurus of Geographic Names (TGN), or the Library of Congress Name Authority File (LCNAF), or to domain-specific or locally developed ontologies. Thus, in this technique, researchers use classifying and clustering algorithms to extract relevant metadata from texts. Term-frequency statistics or IF.IDF, which determines likelihood of keyword applicability through its relative frequency within a given

document as opposed to its relative infrequency in related documents, are commonly used in this technique.

Projects such as John Hopkins University's Automatic Name Authority Control (ANAC) tool utilizes this technique to extract the names of composers within its sheet music collections and to assign the authorized form of those names based on comparisons with LCNAF.[11] Erbs et al. also use this technique to extract key phrases from German educational documents which are then used to assign index terms, thereby increasing the degree to which related documents are collocated within the repository and the consistency of subject term application.[12]

Table 3 lists the tools that support automatic indexing either as the sole technique or as one of a suite of techniques used to generate metadata from resources. Of the thirty-nine tools evaluated for this study, seven tools support some form of automatic indexing.

Tool Name	Location	Techniques	Functions/Features
Apache POI—Text Extractor	http://poi.apache.org/download.html	content extractor; meta-tag harvester; extrinsic auto-generator	Apache POI provides basic text extraction for all project supported file formats. In addition to the (plain) text, Apache POI can access the metadata associated with a given file, such as title and author.
Apache Tika	http://tika.apache.org/	content extractor; meta-tag harvester; extrinsic auto-generator	Built on Apache POI, the Apache Tika toolkit detects and extracts metadata and text content from various documents.
Data Fountains	http://datafountains.ucr.edu/	content extractor; automatic indexer; meta-tag harvester; extrinsic auto-generator	Scans HTML documents and first extracts information contained in meta-tags. If information is unavailable in meta-tags, the program will use other techniques to assign values. Includes a focused web crawler that can target websites concerning a specific subject.
Digital Record Object Identification (DROID)	http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/	extrinsic auto-generator	DROID is a software tool developed by the National Archives to perform automated batch identification of file formats.
Dspace	http://www.dspace.org/	meta-tag harvester; extrinsic auto-generator	Automatically extracts technical information regarding file format and size. Can also extract some information from meta-tags.
Editor-Converter Dublin Core Metadata	http://www.library.kr.ua/dc/dceditunie.html	meta-tag harvester; extrinsic auto-generator	Scans HTML documents, harvesting metadata from tags and converting them to Dublin Core.

doi: 10.6017/ital.v34i3.5889

Embedded Metadata Extraction Tool (EMET)	http://www.artstor.org/global/g-html/download-emet-public.html	content extractor; meta-tag harvester; extrinsic auto-generator	EMET is a tool designed to extract metadata embedded in JPEG and TIFF files.
Firefox Dublin Core Viewer Extension	http://www.splintered.co.uk/experiments/73/	meta-tag harvester; extrinsic auto-generator	Scans HTML documents, harvesting metadata from tags and displaying them to Dublin Core.
JHove	http://jhove.sourceforge.net/#implementation	extrinsic auto-generator	Extracts metadata regarding file format and size as well as validating the structure of the identified file format.
National Library of New Zealand—Metadata Extraction Tool	http://meta-extractor.sourceforge.net/	extrinsic auto-generator	Developed by the National Library of New Zealand to programmatically extract preservation metadata from a range of file formats like PDF documents, image files, sound files, Microsoft Office documents, and others.
Omeka	http://omeka.org/	extrinsic auto-generator; social tagging	Automatically extracts technical information regarding file format and size.
RepoMMan	http://www.hull.ac.uk/esig/repomman/index.html	meta-tag harvester; content extractor; extrinsic auto-generator	Automatically extracts various elements for documents uploaded to Fedora such as author, title, description, and key words, among others. Results are presented to user for review.
Simple Automatic Metadata Generation Interface (SamgI)	http://hmdb.cs.kuleuven.be/amg/Download.php	content extractor; extrinsic auto-generator	A suite of tools that is able to automatically extract metadata elements such as keyphrase and language from documents as well as from the context in which a document exists.

Table 3. Semi-automatic Tools that Support Automatic Indexing

Text and Data Mining

The two methods discussed above, content extraction and automatic indexing, rely on text- and data-mining techniques for the automatic extraction of metadata. In other words, the above methods utilize machine-learning algorithms, statistical analysis of term frequencies, clustering techniques, or techniques that examine the frequency of term utilization between documents as opposed to the use of controlled vocabularies, and classifying techniques, or techniques that exploit the conventional structure of documents, for the semi-automatic generation of metadata. Because of the complexity of these techniques, few tools have been fully developed for application within real-world library settings. Rather, most uses of these techniques have been developed to solve the problems of automatic metadata generation within the context of specific research projects.

There are two reasons for this. One is that, as many researchers have noted, the effectiveness of machine learning techniques depends on the quality and quantity of training data used to teach the system.[13, 14, 15] Because of the number and diversity of subject domains as well as the shear variety of document formats, many applications are designed to address the metadata needs of very specific subject domains and very specific types of documents. This is a point that Kovacevic et al. make in stating that machine learning techniques generally work best for documents of a similar type, like research papers.[16] Another issue, especially as it applies to automatic indexing, is the fact that, as Gardner notes, controlled vocabularies such as the LCSH are too complicated and diverse in structure to be applied through semi-automatic means.[17] Although some open-source tools such as Data Fountains have made efforts to overcome this complexity, projects like it are the exception rather than the rule. These issues signify the difficulty with developing sophisticated semi-automatic metadata generation tools that have general applicability across a wide range of subject domains and format types. Nevertheless, for semi-automatic metadata generation tools to become a reality for the library community, such complexity will have to be overcome.

There are, however, some tools that have broader applicability or can be customized to meet local needs. For instance, the Kea keyphrase extractor offers the option of building local or applying available ontologies that can be used to refine the extraction process. Perhaps the most promising of all is the above mentioned Data Fountains suite of tools developed by the University of California. The Data Fountains suite incorporates almost every one of the semi-automatic metadata techniques described in this study, including sophisticated content extraction and automatic indexing features. It also provides several ways to customize the suite in order to meet local needs.

Extrinsic Data Auto-Generation

Extrinsic data auto-generation is the process of extracting metadata about an information resource that is not contained within the resource itself. Extrinsic data auto-generation can involve the extraction of technical metadata such as file format and size but can also include the extraction of more complicated features such as the grade level of an educational resource or the intended audience for a document. The process of extracting technical metadata is perhaps one area of semi-automatic metadata generation that is in a high state of development, included in most CMSS such as Dspace,[18] as well as other more sophisticated tools such as Harvard's JHove, which can recognize at least 7twelve different kinds of textual, audio, and visual file formats.[19] On the other hand, the problem of semi-automatically generating other types of extrinsic metadata, like grade level, are of the most difficult to solve.

As Leibbrandt et al. note in their analysis of the use of artificial intelligence mechanisms to generate subject metadata for a repository of educational materials at the Education Services Australia, the extraction of extrinsic metadata such as grade level was much more difficult than the extraction of keywords because of the lack of information surrounding a resource's context within the resource itself.[20] This difficulty can also be seen in the absence of tools that support the

doi: 10.6017/ital.v34i3.5889

extraction of extrinsic data beyond those that are harvesting metadata that has been created manually or extracting technical metadata.

Table 4 lists the tools that support extrinsic data auto-generation either as the sole technique or as one of a suite of techniques used to generate metadata from resources. Of the thirty-nine tools evaluated for this study, thirteen tools support some form of extrinsic data auto-generation.

Tool Name	Location	Techniques	Functions/Features
Apache POI—Text Extractor	http://poi.apache.org/download.html	content extractor; meta-tag harvester; extrinsic auto-generator	Apache POI provides basic text extraction for all project supported file formats. In addition to the (plain) text, Apache POI can access the metadata associated with a given file, such as title and author.
Apache Tika	http://tika.apache.org/	content extractor; meta-tag harvester; extrinsic auto-generator	Built on Apache POI, the Apache Tika toolkit detects and extracts metadata and text content from various documents.
Data Fountains	http://datafountains.ucr.edu/	content extractor; automatic indexer; meta-tag harvester; extrinsic auto-generator	Scans HTML documents and first extracts information contained in meta-tags. If information is unavailable in meta-tags, the program will use other techniques to assign values. Includes a focused web crawler that can target websites concerning a specific subject.
Digital Record Object Identification (DROID)	http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/	extrinsic auto-generator	DROID is a software tool developed by the National Archives to perform automated batch identification of file formats.
Dspace	http://www.dspace.org/	meta-tag harvester; extrinsic auto-generator	Automatically extracts technical information regarding file format and size. Can also extract some information from meta-tags.
Editor-Converter Dublin Core Metadata	http://www.library.kr.ua/dc/dceditunie.html	meta-tag harvester; extrinsic auto-generator	Scans HTML documents, harvesting metadata from tags and converting them to Dublin Core.
Embedded Metadata Extraction Tool (EMET)	http://www.artstor.org/global/g-html/download-emet-public.html	content extractor; meta-tag harvester; extrinsic auto-generator	EMET is a tool designed to extract metadata embedded in JPEG and TIFF files.
Firefox Dublin Core Viewer Extension	http://www.splintered.co.uk/experiments/73/	meta-tag harvester; extrinsic auto-generator	Scans HTML documents, harvesting metadata from tags and displaying them to Dublin Core.
JHove	http://jhove.sourceforge.net/	extrinsic auto-	Extracts metadata regarding file

	#implementation	generator	format and size as well as validating the structure of the identified file format.
National Library of New Zealand—Metadata Extraction Tool	http://meta-extractor.sourceforge.net/	extrinsic auto-generator	Developed by the National Library of New Zealand to programmatically extract preservation metadata from a range of file formats like PDF documents, image files, sound files, Microsoft Office documents, and others.
Omeka	http://omeka.org/	extrinsic auto-generator; social tagging	Automatically extracts technical information regarding file format and size.
RepoMMan	http://www.hull.ac.uk/esig/repomman/index.html	meta-tag harvester; content extractor; extrinsic auto-generator	Automatically extracts various elements for documents uploaded to Fedora such as author, title, description, and key words, among others. Results are presented to user for review.
Simple Automatic Metadata Generation Interface (SamgI)	http://hmdb.cs.kuleuven.be/amg/Download.php	content extractor; extrinsic auto-generator	A suite of tools that is able to automatically extract metadata elements such as keyphrase and language from documents as well as from the context in which a document exists.

Table 4. Semi-Automatic Tools that Support Extrinsic Data Auto-Generation.

Social Tagging

Social tagging is now a familiar form of subject metadata generation although, as mentioned previously, it is not properly a form of automatic metadata generation. Nevertheless, because of the relatively low cost in generating and maintaining metadata through social tagging and its current widespread popularity, a few projects have attempted to utilize such data to enhance repositories. For instance, Linstaedt et al. use sophisticated computer programs to analyze still images found within Flickr and then use this analysis to process new images and to propagate relevant user tags to those images.[21]

In a slightly more complicated example, Liu and Qin employ machine-learning techniques to initially process and assign metadata, including subject terms, to a repository of documents related to the computer science profession.[22] However, this proof of concept project also permits users to edit the fields of the metadata once established. The user-edited tags are then reprocessed by the system with the hope of improving the machine-learning mechanisms of the database, creating a kind of feedback loop for the system. Specifically, the improved tags are used by the system to suggest and assign subject terms for new documents as well as to improve subject description of existing documents within the repository. Although these two examples provide instances of sophisticated reprocessing of social tag metadata, these capabilities do not seem to be present in open-source tools at this time. Nevertheless, social tagging capabilities are offered by many CMSS such as Omeka. These social tagging capabilities may offer a means to enhance subject access to holdings.

doi: 10.6017/ital.v34i3.5889

Table 5 below lists the tools that support social tagging either as the sole technique or as one of a suite of techniques used to generate metadata from resources. Of the thirty-nine tools evaluated for this study, two tools support some form of social tagging.

Tool Name	Location	Techniques	Functions/Features
Dspace	http://www.dspace.org/	meta-tag harvester; extrinsic auto-generator; social tagging	Automatically extracts technical information regarding file format and size. Can also extract some information from meta-tags.
Omeka	http://omeka.org/	extrinsic auto-generator; social tagging	Automatically extracts technical information regarding file format and size.

Table 5. Semi-automatic Tools that Support Social Tagging.

Challenges to Implementation

Although semi-automatic metadata generation tools offer many benefits, especially in regards to streamlining the metadata-creation process, there are significant barriers to the widespread adoption and implementation of these tools. One problem with semi-automatic metadata generation tools is that many are developed locally to address the specific needs of a given project or as part of academic research. This local, highly focused milieu for development means that general applicability of the tools is potentially diminished. The local context may also hinder widespread adoption of applications that would result in strong communities of application users and provide further support for the development of applications in an open-source context. Because of the highly specific nature of many current tools, their relevance to real-world processes of metadata creation within the broader context of libraries' diverse information management needs are not accounted for.

Additionally, many tools are focused on solving one or, at most, a few metadata generation problems. For instance, the Kea application is designed to use machine-learning techniques for the sole purpose of extracting keywords, the Open Text Summarizer is limited to automatic extractions of summary descriptions and keywords, and Editor Converter Dublin Core is designed to extract information in HTML meta-tags and map them to Dublin Core elements. Because of the piecemeal development of semi-automatic generation tools, any comprehensive package of tools will require the significant efforts of the implementer to coordinate the selected applications and to produce results in a single output. This is, to say the least, a daunting task.

Furthermore, a high degree of technical skill is required to implement these complex tools. Many of the more sophisticated tools used to semi-automatically generate metadata, such as Data Fountains, Kea, and Apache Stanbol, require competence in a variety of programming languages.

Significant knowledge of C++, Python, and Java, are required to implement these systems properly. The high degree of technical knowledge needed to implement these tools means that many libraries and other institutions may not have resources to begin implementing them, let alone incorporating them into the daily workflows of the metadata creation process. Further, this high degree of technical expertise may require libraries to seek assistance outside of the library. In other words, librarians may need to build strong collaborative relationships with those who have the technical skills, expertise and credentials to implement and maintain these complicated tools. As Vellucci et al. note in regards to their development of the Metadata Education and Research Information Commons (MERIC), a metadata-driven clearinghouse of education materials related to metadata, elaborate and multidisciplinary partnerships need to be firmly established for the ultimate success of such projects, including the sustained support of the highest levels of administration.[23] These types of partnerships may be difficult to establish and maintain for the sustained implementation of complicated tools.

Additionally, sustainable development of tools, especially in regards to the funding needed for continued development of open-source applications, appears to be a significant barrier to implementation. For instance, at the time of this writing, many of the tools that were touted in the literature as being most promising, such as DC Dot, Reggie, and DescribeThis, are no longer available for implementation. Beyond the fact that discontinuation hurts the potential adoption and continued development of semi-automatic tools within real world library and other information settings, there is also the problem that those settings that have in fact adopted tools may lose the technical support of a central developer and community of users. Thus discontinuation may result in higher rates of tool obsolescence and increase the potential expenses of libraries who have implemented and then must change applications.

Finally, the application of semi-automatic metadata tools remains relatively untested in real-world scenarios. As Polfreman et al. note, most tests of automatic metadata generation tools have several of problems, including small sample sizes, narrow scope of project domains, and experiments that lack true objectivity because systems are generally tested by their creators.[24] For these reasons, libraries and other institutions may be reluctant to expand the resources needed to implement and fully integrate a complicated, promising, but ultimately untested, tool within the already strained workflows of its processes.

CONCLUSION

Semi-automatic metadata generation tools hold the promise of assisting information professionals with the management of ever-increasing quantities and types of information resources. Using software that can create metadata records consistently and efficiently, semi-automatic metadata generation tools potentially offer significant cost and time savings. However, the full integration of these tools into the daily workflows of libraries and other information settings remains elusive. For instance, although many tools have been developed that have addressed many of the more complicated aspects of semi-automatic metadata generation, including the extraction of

doi: 10.6017/ital.v34i3.5889

information related to conceptually difficult areas of bibliographic description such as subject terms, open-ended resource descriptions, and keyword assignment, many of these tools are relevant only at the project level and are not applicable to the broader contexts needed by libraries. In other words, the current array of tools exists to solve experimental problems but has not been developed to the point that the library community can implement it in a meaningful way.

Perhaps the greatest area of difficulty lies in the fact that most tools only address part of the problem of semi-automatic metadata generation, providing solutions to the semi-automatic generation of one or a few bibliographic elements but not the full range elements. This means that for libraries to truly have a comprehensive tool set for the semi-automatic generation of metadata records, significant local efforts will be required to integrate the various tools into a working whole. Couple this issue with the instability of tool development and maintenance and it appears that the library community may lack incentive to invest already strained and limited resources in the adoption of these tools.

Thus it appears that a number of steps will need to be taken before the library community can seriously consider the incorporation of semi-automatic metadata generation tools within its daily workflows. First, it seems that the integration of these various tools into a coherent set of applications is likely the next step in the development of viable semi-automatic metadata generation. Since most small libraries likely do not have the resources required to integrate these disparate tools together, let alone incorporate them within existing library systems, a single package of tools will be needed simply from a resource perspective. Secondly, considering the high level of technical expertise needed to implement the current array of tools, the integrated set of tools must be accomplished in such a way as to foster implementation, utilization, and maintenance with a minimum of technical expertise. For instance, if an integrated set of tools that functioned across a wide range of subject domains and format types could be developed, the suite might be akin to the CMSS currently employed by many libraries. Furthermore, with a suite of tools that are relatively easy to use, adaption would likely increase. This might result in a stable community of users that would foster the further development of the tools in a sustainable manner. A comprehensive, relatively easy to implement set of tools might foster independent testing of those tools. The independent testing of the semi-automatic tools is needed to provide an objective basis for tool evaluation and further development.

Finally, designing automated workflows tailored to the subject domain and types of resources seems to be an essential step for integrating semi-automatic metadata generation tools into metadata creation. Such workflows may delineate data elements that can be generated by automated meta-tag extractor from data elements that need to be refined and manually created by cataloging and metadata professionals. To develop, maximize, and sustain semi-automatic metadata generation workflows, administrative support for finance, human resources, and training is critical.

Thus, although many of the technical aspects of semi-automatic metadata generation are well on their way to being solved, many other barriers exist that might limit adoption. Further, these barriers may have a negative influence on the continued, sustainable development of semi-automatic metadata generation tools. Nevertheless, there is a critical need that the library community finds ways to manage the recent explosion of data and information in cost-effective and efficient ways. Semi-automatic metadata generation holds the promise to do just that.

ACKNOWLEDGEMENT

This study was supported by the Institute of Museum and Library Services.

REFERENCES

1. Jane Greenberg, Kristina Spurgin, and Abe Crystal, "Final Report for the AMeGA (AutoZmatic
2. Sue Ann Gardner, "Cresting Toward the Sea Change," *Library Resources & Technical Services* 56, no. 2 (2012): 64–79, http://dx.doi.org/10.5860/lrts.56n2.64.
3. For details, see Jung-ran Park and Caimei Lu, "Application of Semi-Automatic Metadata Generation in Libraries: Types, Tools, and Techniques," *Library & Information Science Research* 31, no. 4 (2009): 225–31, http://dx.doi.org/10.1016/j.lisr.2009.05.002.
4. Erik Mitchell, "Trending Tech Services: Programmatic Tools and the Implications of Automation in the Next Generation of Metadata," *Technical Services Quarterly* 30, no. 3 (2013): 296–10, http://dx.doi.org/10.1080/07317131.2013.785802.
5. Jane Greenberg, "Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications," *Journal of Internet Cataloging* 6, no. 4 (2004): 59–82, http://dx.doi.org/10.1300/J141v06n04_05.
6. Malcolm Polfreman, Vanda Broughton, and Andrew Wilson, "Metadata Generation for Resource Discovery," JISC, 2008, http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/autometgen.aspx.
7 Park and Lu, "Application of Semi-Automatic Metadata Generation in Libraries."
8. Kea Automatic Keyphrase Extraction homepage, http://www.nzdl.org/Kea/index_old.html.
9. Wilhelmina Randtke, "Automated Metadata Creation: Possibilities and Pitfalls," *Serials Librarian* 64, no. 1–4 (2013): 267–84, http://dx.doi.org/10.1080/0361526X.2013.760286.
10. Aleksandar Kovačević et al.,"Automatic Extraction of Metadata from Scientific Publications for CRIS Systems." *Electronic Library and Information Systems* 45, no. 4 (2011): 376–96, http://dx.doi.org/10.1108/00330331111182094.

11. Mark Patton et al., “Toward a Metadata Generation Framework: A Case Study at Johns Hopkins University,” *D-Lib Magazine* 10, no. 11 (2004), http://www.dlib.org/dlib/november04/choudhury/11choudhury.html.

12. Nicolai Erbs, Iryna Gurevych, and Marc Rittberger, “Bringing Order to Digital Libraries: From Keyphrase Extraction to Index Term Assignment.” *D-Lib Magazine* 19, no. 9/10 (2013), http://www.dlib.org/dlib/september13/erbs/09erbs.html.

13. Polfreman, Broughton, and Wilson, “Metadata Generation for Resource Discovery.”

14. Randtke, “Automated Metadata Creation.”

15. Xiaozhong Liu and Jian Qin, “An Interactive Metadata Model for Structural, Descriptive, and Referential Representation of Scholarly Output,” *Journal of the Association for Information Science & Technology* 65, no. 5 (2014): 964–83, http://dx.doi.org/10.1002/asi.23007.

16. Kovačević et al., “Automatic Extraction of Metadata from Scientific Publications for CRIS Systems.”

17. Gardner, “Cresting Toward the Sea Change.”

18. Mary Kurtz, “Dublin Core, Dspace, and a Brief Analysis of Three University Repositories,” *Information Technology & Libraries* 29, no. 1 (2010): 40–46, http://dx.doi.org/10.6017/ital.v29i1.3157.

19. “JHOVE - JSTOR/Harvard Object Validation Environment,” JSTOR, http://jhove.sourceforge.net.

20. Richard Leibbrandt et al., “Smart Collections: Can Artificial Intelligence Tools and Techniques Assist with Discovering, Evaluating and Tagging Digital Learning Resources?” *International Association of School Librarianship: Selected Papers from the Annual Conference* (2010).

21. Stefanie Lindstaedt et al., “Automatic Image Annotation Using Visual Content and Folksonomies,” *Multimedia Tools & Applications* 42, no. 1 (2009): 97–113, http://dx.doi.org/10.1007/s11042-008-0247-7.

22. Liu and Qin, “An Interactive Metadata Model.”

23. Sherry Vellucci, Ingrid Hsieh-Yee, and William Moen, “The Metadata Education and Research Information Commons (MERIC): A Collaborative Teaching and Research Initiative,” *Education for Information* 25, no. 3/4 (2007): 169–78.

24. Polfreman, Broughton, and Wilson, “Metadata Generation for Resource Discovery.”